Astronomy cross-matching using OGSA-DAI

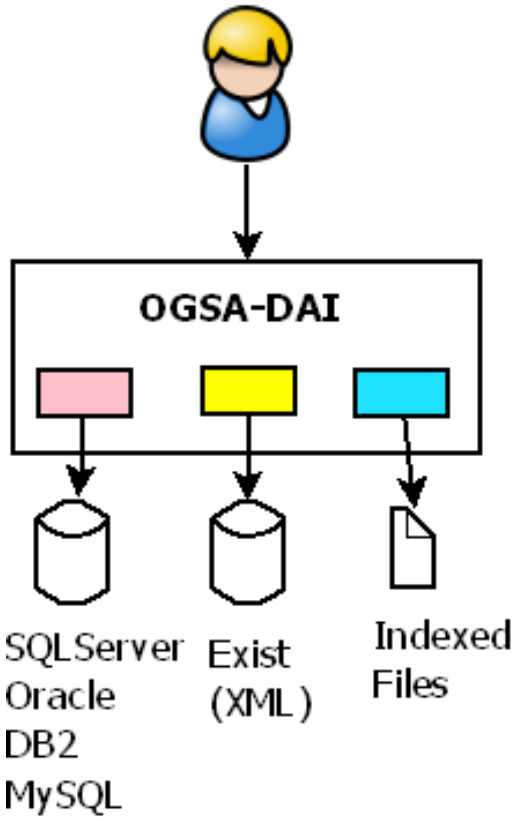
Keith Noddle, WFAU Tech Lead

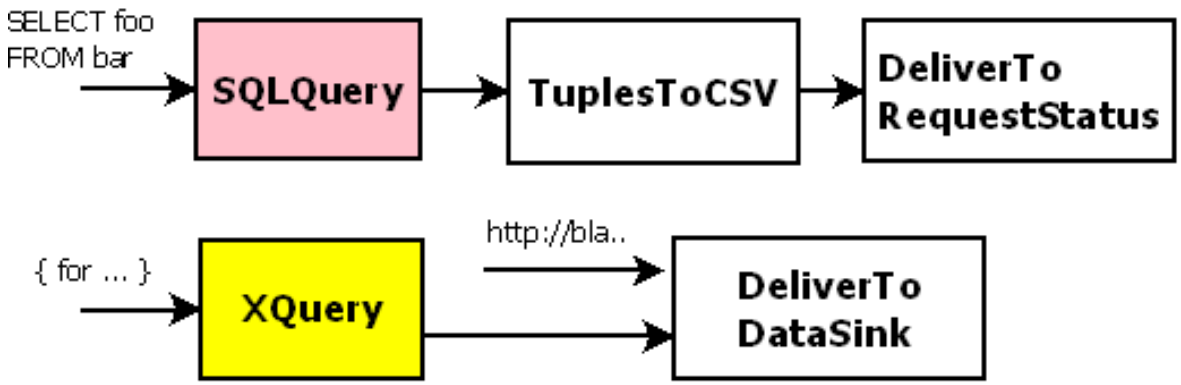
originally presented by:
Ally Hume, EPCC, The University of Edinburgh

http://www.ogsadai.org.uk

ktn@roe.ac.uk / A.Hume@epcc.ed.ac.uk

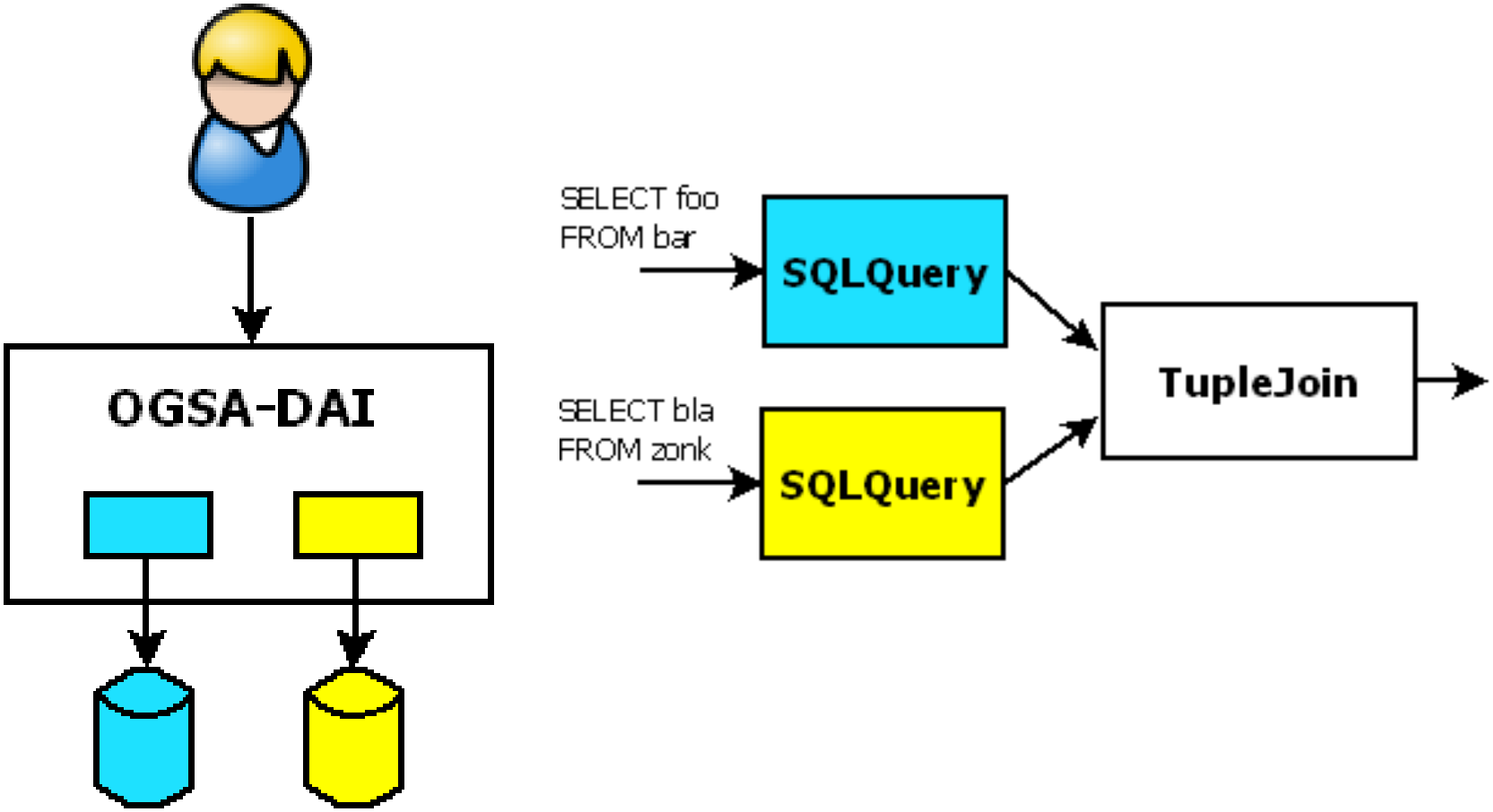
Overview



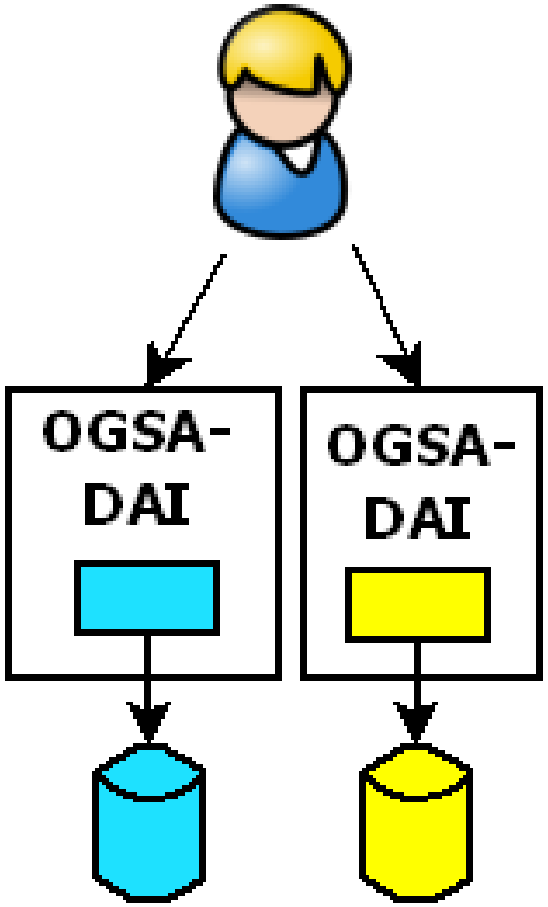


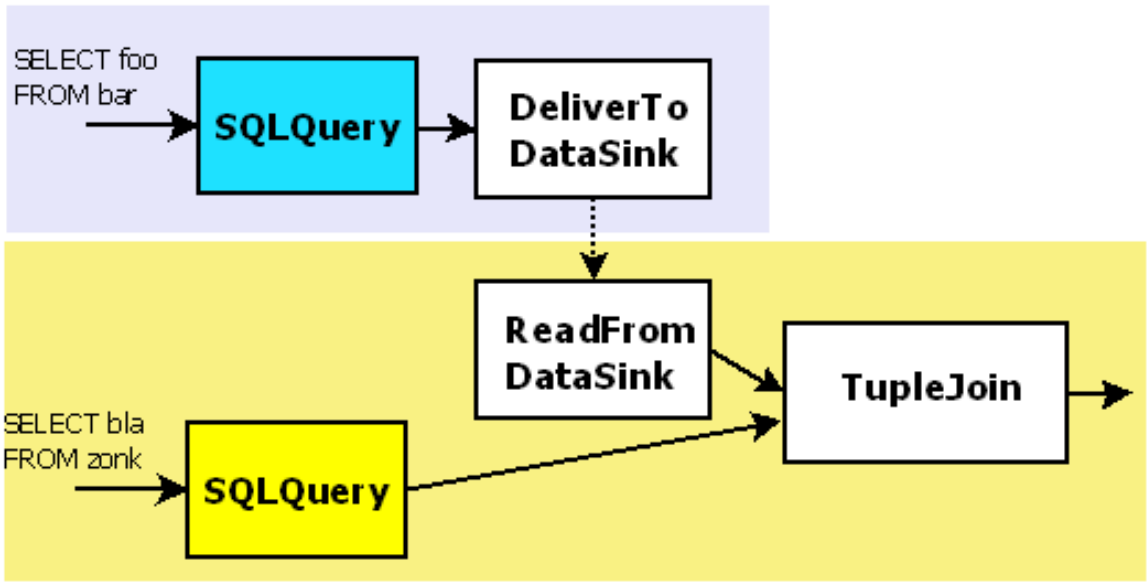
ktn@roe.ac.uk / A.Hume@epcc.ed.ac.uk

OGSA-DAI joins



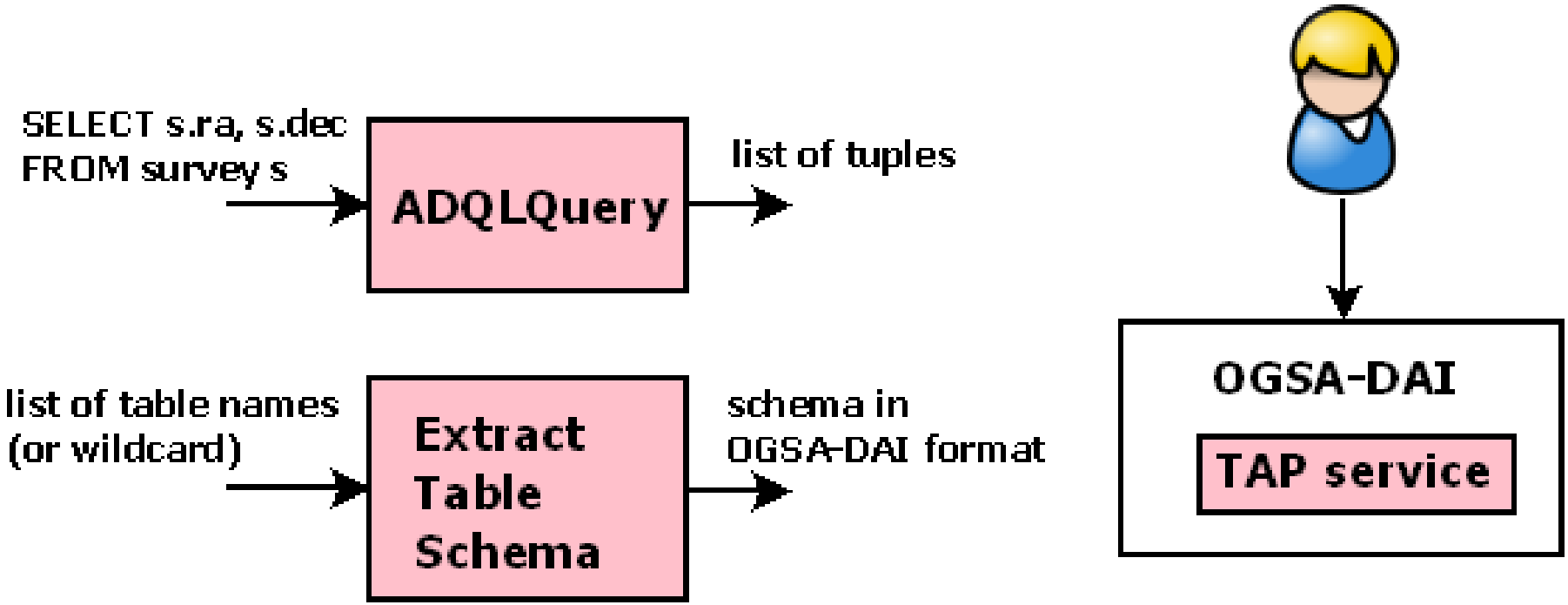
OGSA-DAI distributed joins





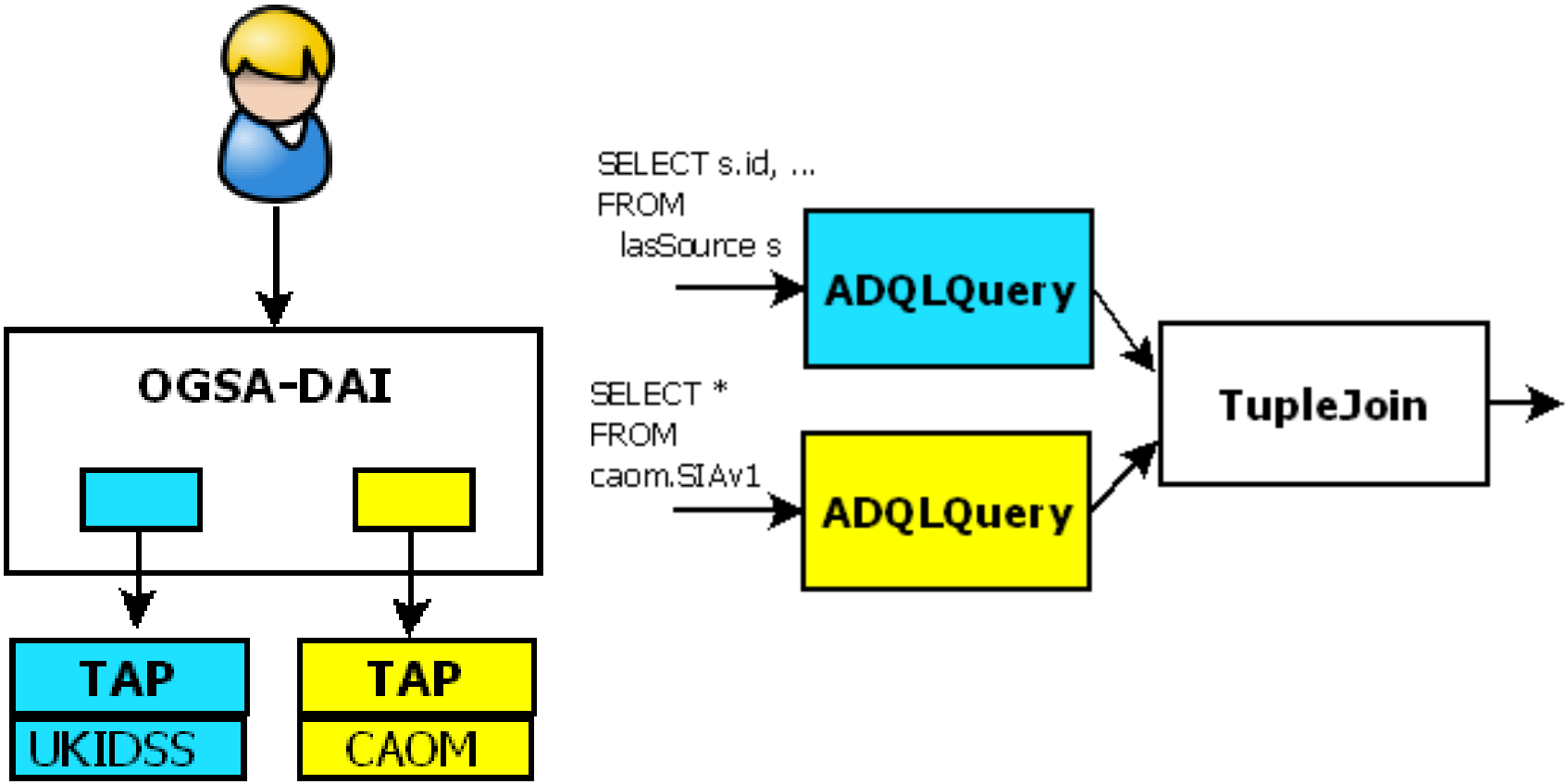
ktn@roe.ac.uk / A.Hume@epcc.ed.ac.uk

OGSA-DAI activities for astronomy



ktn@roe.ac.uk / A.Hume@epcc.ed.ac.uk

Astronomy cross-match

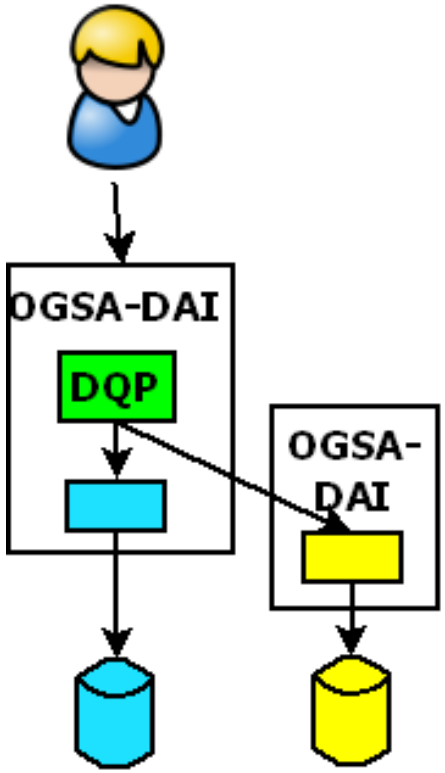


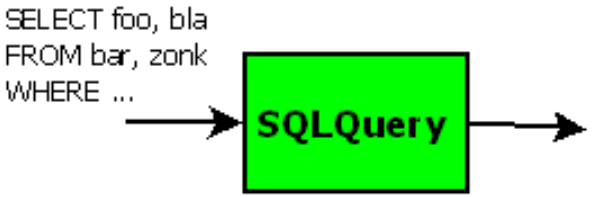
ktn@roe.ac.uk / A.Hume@epcc.ed.ac.uk

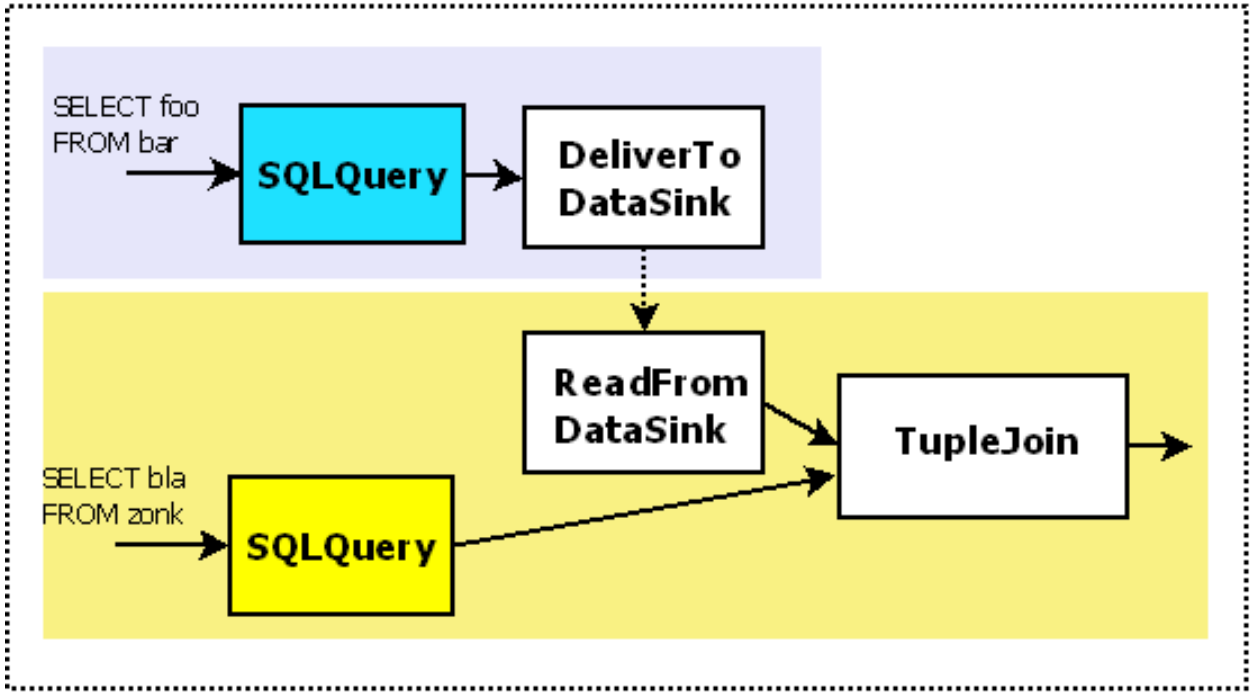
OGSA-DAI join options

- In-memory join
  - One side of join stored in memory, other side streamed
- Partial in-memory join
  - Gets first results quickly, but all data stored to disk
- Ordered merge join
  - Both inputs ordered allowing for a fully streamed join
- Parallel hash equi-join
- Batch joins using IN clauses
  - E.g. SELECT * FROM foo WHERE bar IN (x, y, z)

Distributed Query Processing (DQP)

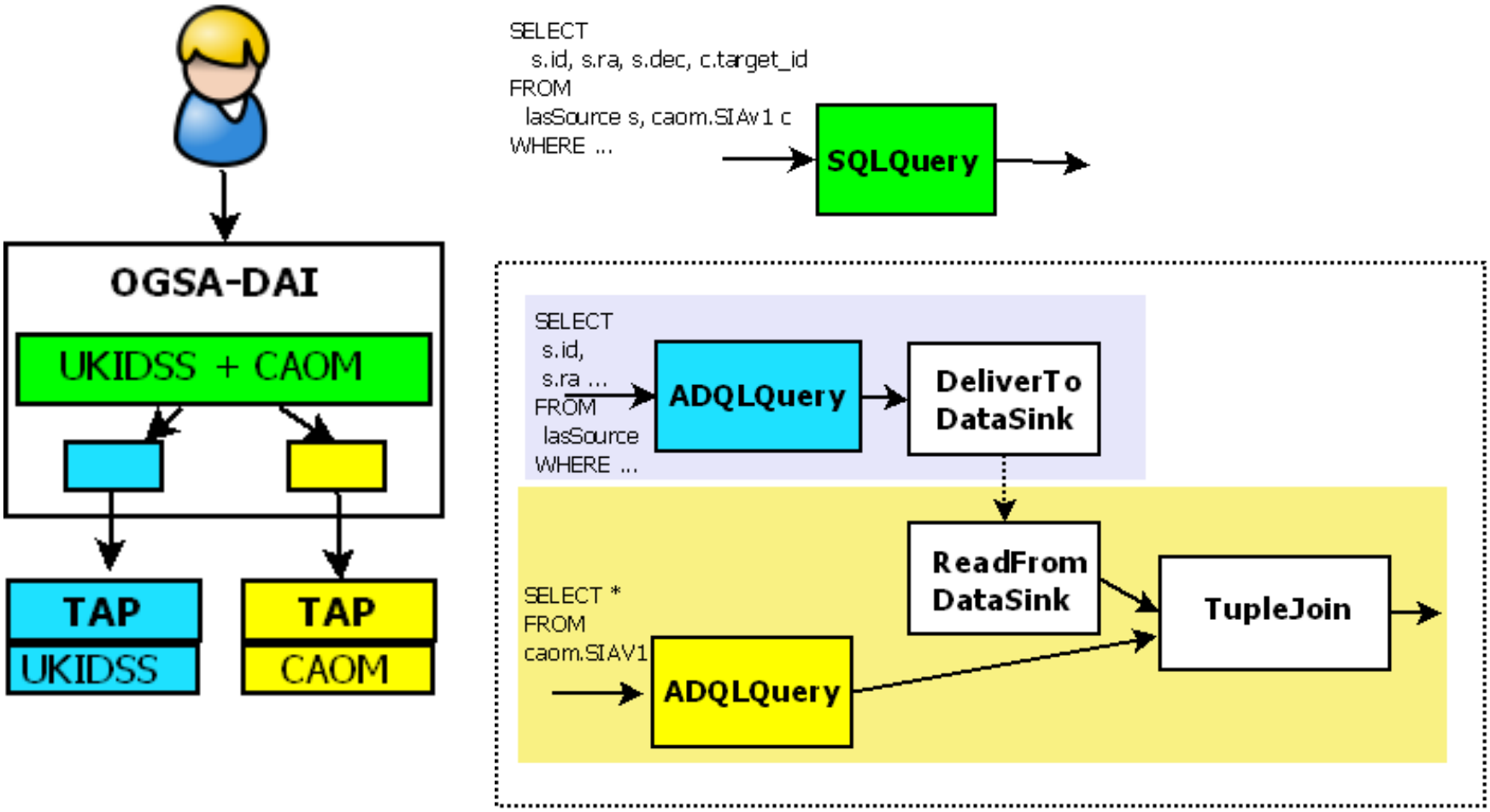






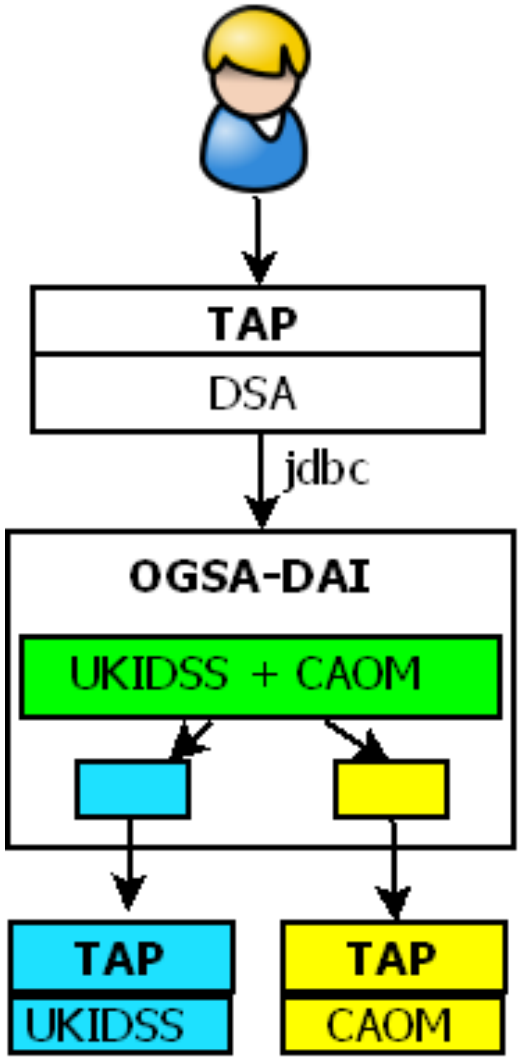
ktn@roe.ac.uk / A.Hume@epcc.ed.ac.uk

DQP over TAP



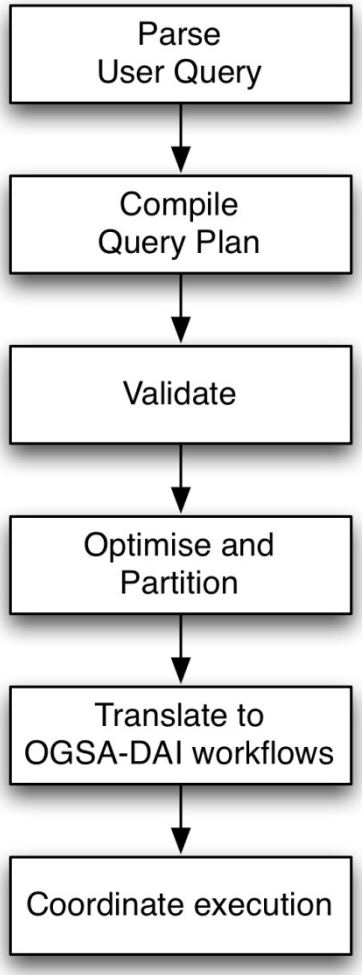
ktn@roe.ac.uk / A.Hume@epcc.ed.ac.uk

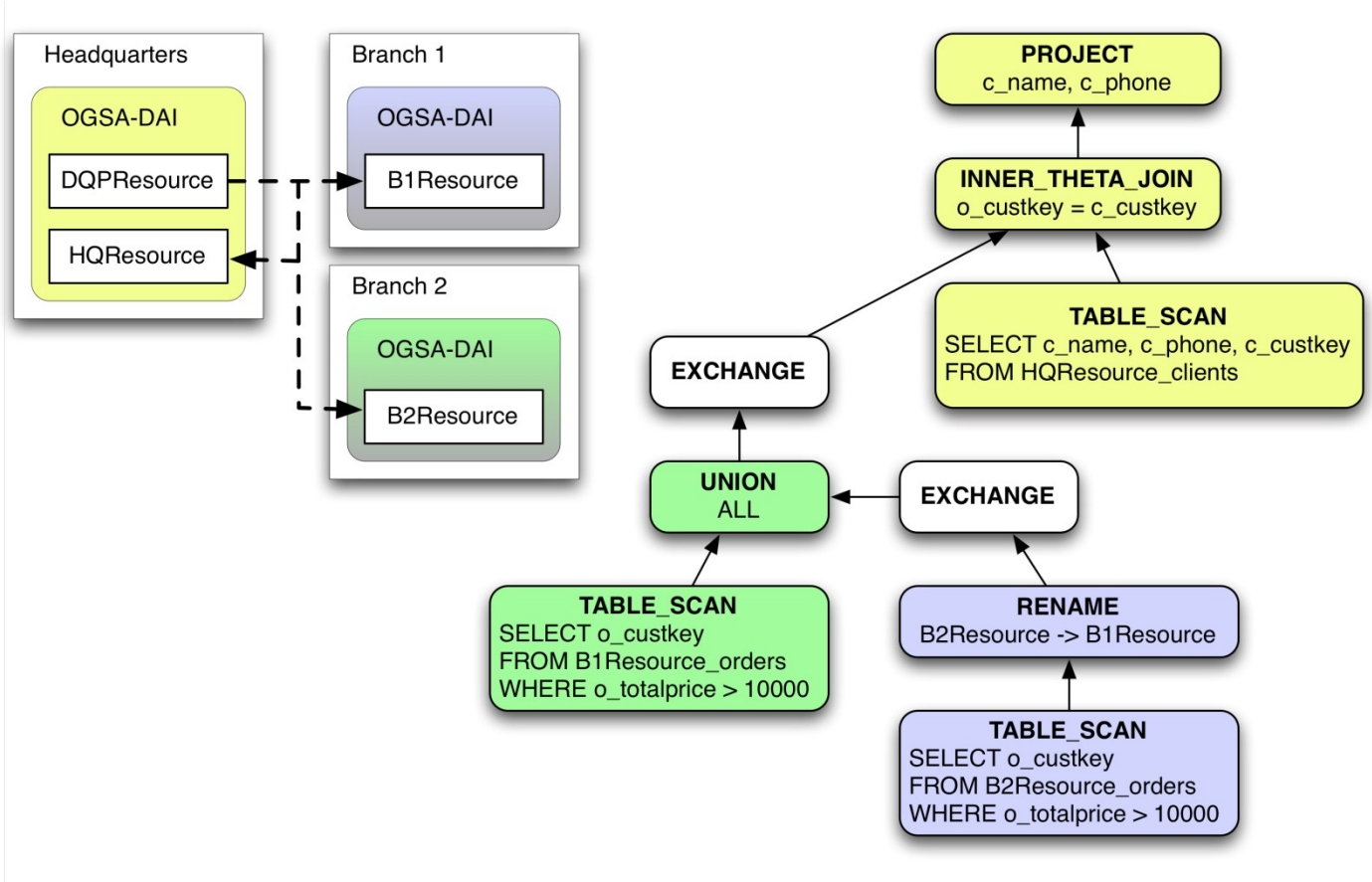
TAP over DQP over TAP



ktn@roe.ac.uk / A.Hume@epcc.ed.ac.uk

DQP under the covers





ktn@roe.ac.uk / A.Hume@epcc.ed.ac.uk

Current status

- Working on two demo queries
- UKIDSS and CAOM
  - In-memory join
  - Works TAP->DQP->TAP for 1 TAP at a time
  - Combined not yet working
- UKIDSS and SDSS
  - Ordered merge join
  - Works direct to DBs with ugly query
  - DQP join ordering needs improved
  - DSA TAP server struggling with the query DQP produces
    - LEFT OUTER JOIN

On-line demos: http://www.ogsadai.org.uk/demos/aida/

UKIDSS/CAOM query


```
select collection, collectionID, target_name, UKIDSS_lasSource.ra,
    UKIDSS_lasSource.dec,UKIDSS_lasSource.yapermag3,

     UKIDSS_lasSource.j_1apermag3,UKIDSS_lasSource.hapermag3,UKIDSS_lasSource.kaperm
     ag3
FROM UKIDSS_lasSource, CAOM_caom_SIAv1
WHERE
   (UKIDSS_lasSource.priOrSec <= 0 OR
    UKIDSS_lasSource.priOrSec = UKIDSS_lasSource.frameSetID)  AND
   UKIDSS_lasSource.yClass >= -2 AND UKIDSS_lasSource.yClass <= -1 AND
   UKIDSS_lasSource.yppErrBits < 256 AND    UKIDSS_lasSource.j_1Class >= -2 AND
   UKIDSS_lasSource.j_1Class <=  -1 AND UKIDSS_lasSource.j_1ppErrBits < 256  AND
   UKIDSS_lasSource.hClass >= -2 AND UKIDSS_lasSource.hClass <= -1 AND
   UKIDSS_lasSource.hppErrBits < 256 AND    UKIDSS_lasSource.kClass >= -2 AND
   UKIDSS_lasSource.kClass <= -1 AND UKIDSS_lasSource.kppErrBits < 256 AND
  (UKIDSS_lasSource.j_1apermag3 - UKIDSS_lasSource.kapermag3) > 1.1 AND
   UKIDSS_lasSource.j_1apermag3 < 18.0 AND
   UKIDSS_lasSource.ra - position_center_ra > -0.1 AND
   UKIDSS_lasSource.ra - position_center_dec < 0.1 AND
   UKIDSS_lasSource.dec - position_center_dec > -0.1 AND
   UKIDSS_lasSource.dec - position_center_dec < 0.1
```

Future

- More join algorithms
  - Batch join into DQP
- Better join choices
  - Need to improve cardinality estimates
    - Need attribute histograms
- Support for ADQL spatial functionality
  - Possibly also special support for spatial joins using bounding boxes
- OGSA-DAI extensions
  - Correlated sub-queries
  - Better scalability for astronomy sided databases where as much as possible must be streamed
  - Missing functionality e.g. TOP

Thanks to...

ktn@roe.ac.uk / A.Hume@epcc.ed.ac.uk

epcc